B.3. Recuperación de información, procesamiento de lenguaje natural y web semántica

Por **José Ramón Pérez Agüera**

Pérez Agüera, José Ramón. "Recuperación de información, procesamiento de lenguaje natural y web semántica". En: *Anuario ThinkEPI*, 2007, pp. 69-70.

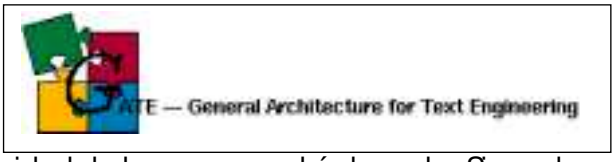
"Animo a los documentalistas a que investiguen en la Web semántica y buceen en la literatura de PLN y RI que ya existe y que es fundamental para hacer realidad el sueño de Tim Berners-Lee"

"Aprender a programar es la única forma de hacer que la Web semántica sea una Web más bibliotecaria, y no solamente más informática"

"La generación de contenidos semánticos no es viable de forma manual, por lo que es necesario la automatización de las tareas"

UNO DE LOS TÉRMINOS MÁS UTILIZADOS cuando nos referimos a la Web es Web semántica. "*Como si de Harry Potter se tratara, todo experto en internet que se precie, debe conocer estas dos palabras mágicas símbolo del futuro de una internet más ordenada, más organizada, más catalogada, en definitiva más bibliotecaria*" (Eva Méndez).

No es mi objetivo definir ni discutir el significado del término Web semántica, ya que insignes científicos y pensadores se han encargado de hacerlo con mucha mayor claridad de lo que yo podría hacerlo. Sin embargo, como persona pragmática que soy, sí voy a entrar a comentar en esta breve nota cuáles son los elementos, herramientas y utilidades que son necesarios para que la Web semántica deje de ser un concepto y se convierta en una realidad. Por supuesto, toda originalidad en mis planteamientos queda descartada, y no pretendo más que convertirme en un mero compilador de evidencias existentes en el panorama investigador.



En el último congreso internacional sobre Web semántica (Osaka, 18-21 oct. 2005), la presencia de aplicaciones centradas en procesamiento de lenguaje natural (PLN) fue más que notable. De hecho, *Gate*, una conocida aplicación para ingeniería lingüística diseñada en la *Universidad de Sheffield*, fue una de las estrellas invitadas (de forma no oficial) al figurar referenciada en un buen número de los trabajos presentados al congreso.

Ahora bien, la utilidad del PLN para la implementación de una Web más semántica, no es un descubrimiento de este año. En 2004, **Ricardo Baeza-Yates** firmó un interesante trabajo sobre la aplicación de técnicas de PLN a la recuperación de información (RI) donde proponía la Web semántica como una de las principales aplicaciones prácticas de técnicas convidadas de PLN y RI.

Sin duda deben existir bastantes más referencias a la vinculación entre PLN, RI y Web

semántica que ahora mismo se me escapan. Pero mi objetivo aquí no es presentar una relación exhaustiva de la vinculación entre estos tres elementos, sino, más bien, y continuando con mi proselitismo tecnológico en el área de ByD, animar a todos aquellos atrevidos documentalistas que se introducen en la Web semántica como área de investigación, a bucear en la literatura de PLN y RI que ya existe y que es fundamental para hacer realidad el sueño de **Tim Berners-Lee**. Animarles también a manejar las herramientas necesarias para implementar la Web semántica, tales como *Gate*, *Protégé*, *Lucene*, *Nutch*, o tantas otras, aunque para ello haya que aprender a programar, ya que ésta es la única forma de hacer que, realmente, la Web semántica sea una Web más bibliotecaria, y no solamente más informática.

Muchos pueden argumentar que el propio padre de la idea defiende que la Web semántica no es una Web basada en técnicas pertenecientes al área de inteligencia artificial (IA), pero esto no significa que nos podamos utilizar estas técnicas como base y apoyo para la implementación de su idea, ya que, más allá de rencillas académicas, todo lo que nos ayude a hacer realidad una nueva Web es útil independientemente de conceptualizaciones de carácter teórico.

Hay que tener en cuenta que, hoy por hoy, la Web semántica no existe como tal, más allá de implementaciones puntuales de tipo experimental. El hecho de que exista pasa inexorablemente por la generación de contenidos web semánticos que den cuerpo a la idea de una web más organizada. La generación de contenidos de carácter semántico no es asimilable de forma manual por lo usuarios y autores de la Web, por lo que es necesario la automatización de todas, o por lo menos parte de las tareas de generación de contenidos web semánticos. Es aquí donde el PLN y la RI tienen mucho que aportar, ya que permiten la implementación de aplicaciones capaces de generar información de tipo semántico que dote de cuerpo a la Web semántica y la conviertan en una realidad.

Analizadores sintácticos, que permitan comprender la estructura de las frases de forma automática, etiquetadores léxicos, reconocedores de entidades como nombres, fechas lugares, todas ellas son herramientas automáticas esenciales para la generación de contenidos web semánticos. Es más, me atrevo a decir que sin ellas no es posible una web semántica real, ya que el coste de elaboración manual de contenidos semánticos no es asimilable desde ningún punto de vista.

Lo aquí expuesto no invalida ni mucho menos otras ideas sobre la implementación de la Web semántica, pero sí se acerca, o al menos ese era el objetivo, a lo que supone la implementación real de la idea, lo cual es desde mi punto de vista la mejor forma de acallar a aquellos que opinan que la Web semántica es un concepto vacío, una entelequia sin sentido o una utopía irrealizable.

Referencias interesantes:

La profusión con la que se ha utilizado Gate en la ISWC 2005 es una muestra del uso y la aplicación del PLN en la Web semántica.
–http://gate.ac.uk/conferences/iswc2003/
–http://gate.ac.uk/semweb.html
–http://www.cc.gatech.edu/ccg/iswc05/
–Uso de lenguajes documentales en la web semántica
http://www.sedic.es/gt_normalizacion_web-semantica05.htm

***José Ramón Pérez Agüera**, Depto. de Sistemas Informáticos y Programación, Facultad de Informática, Universidad Complutense de Madrid*
jose.aguera@fdi.ucm.es
http://multidoc.rediris.es/joseramon